

# NIST Proof-of-Concept (POC) Exercises for the GALE MT Post Editing

Mark Przybocki, John Garofolo,  
Greg Sanders, Audrey Le

NIST

July 7<sup>th</sup>, 2005

# Objective

- Short Term
  - Test the process of post editing on a meaningful scale on GALE domains of interest
  - Investigate GALE Post-Editing-based protocols and metric(s)
    - Adjudication of independent human references into a gold standard
    - Implementing post editing that includes “production-type” editors
  - Examine inter-editor agreement
  - Identify possible short-comings early on
  - Determine required test-set size to attain the necessary level of statistical significance in comparing systems
- Long Term
  - Implement a dry run evaluation which is completely representative of the formal evaluation in terms of scale and protocols
  - Conduct first formal evaluation

# Schedule

- POC1 – *completed (in less than two weeks)*
  - Arabic Text from MT04
- POC2 – (planned for July)
  - Chinese Text from MT04
- ▶ POC3 – (planned for Aug/Sept)
  - Arabic Speech, speech input (requiring ASR)
  - First post editing exercise using speech input
- Dry run (planned for January 2006)
  - All four GALE input domains, Arabic and Chinese speech and text input
- Formal evaluation (planned for June 2006)
  - Arabic and Chinese speech and text input

# POC1 - Data from MT04

- Documents
  - 10 MT 2004 documents
  - Ranked docs by average BLEU, chose docs at 5<sup>th</sup>, 15<sup>th</sup>, ..., 95<sup>th</sup> percentile
  - 81 segments, 2080 reference words
- Reference
  - NIST adjudicated the four MT04 references into one Gold-Standard
    - Majority rules technique
    - A native Arabic speaker helped to resolve questions and ties
    - Much stricter process needed for the formal dry run and evaluation
      - examine all conflicts
- System output to be Post Edited
  - Three MT04 systems were selected to represent varied performance
    - ISI – 47%, IBM – 34%, FCT – 28% (*MT04 BLEU, 4 – refs, 100 docs*)
    - ISI – 31%, IBM – 20%, FCT – 17% (*This 10doc set BLEU, 1 – GS-ref*)

# POC1 - Guidelines

- Guidelines
  - Modified to reflect feedback from the MT05 workshop

Make the MT output have the *correct meaning*, be *readily understandable*, and really be *English*.

- (1) The goal is to edit the MT output so that it has **the same meaning** as the human translation.
- (2) Edit the MT output so that it is **understandable**.
- (3) Punctuation must be understandable, and sentence-like units must have sentence-ending punctuation. But do not insert, delete, or change other punctuation merely to follow optional traditional rules about what is “strictly correct.”
- (4) If words/phrases/punctuation in the MT output or the reference human translation are completely acceptable, prefer them over substitutions.
- (5) Dates, as well as the commas and decimal points in numbers, should be formatted according to U.S. conventions (for example, convert 23-2-2004 to 2-23-2004).

# POC1 - Guidelines *(cont'd)*

## Included a description of 7 categories of fluency

- (1) Incomprehensible mess
- (2) Ugly but understandable
- (3) Non-native English (includes fundamental errors)
- (4) Communication through speech between two 11 yr. old native English speakers (no syntax or tense errors). The speaker knows English but is not necessarily polished. Casual conversation between adult native speakers of English.
- (5) Spoken English used by a high-school teacher (reasonably educated adult). Typical newspaper news article. Typical prepared business talks.
- (6) Written English that is not a first draft
- (7) English that has been re-written repeatedly and polished to the highest degree by someone with a well-developed sense of style

**Goal of Editor:** Turn MT output into category (4) or (5)  
While reference might be a (5) or (6)

# Post Editor Completion Rate

***updated slide 30-Aug-2005***

Post Editor	Ver*	% Completed	<i>Approximate</i>	
			Time spent editing	Words Per Hour
Doddington	v1	100%	<b>8hr 45min</b>	<b>810</b>
Wayne	v1	100%	<b>9hr 25min</b>	<b>750</b>
Buckland (NIST)	v3	100%	<b>8hr 30min</b>	<b>830</b>
Martin (NIST)	v3	100%	<b>13hr 20min</b>	<b>530</b>
High School Teacher	v2	100%	<b>9hr 05min</b>	<b>785</b>

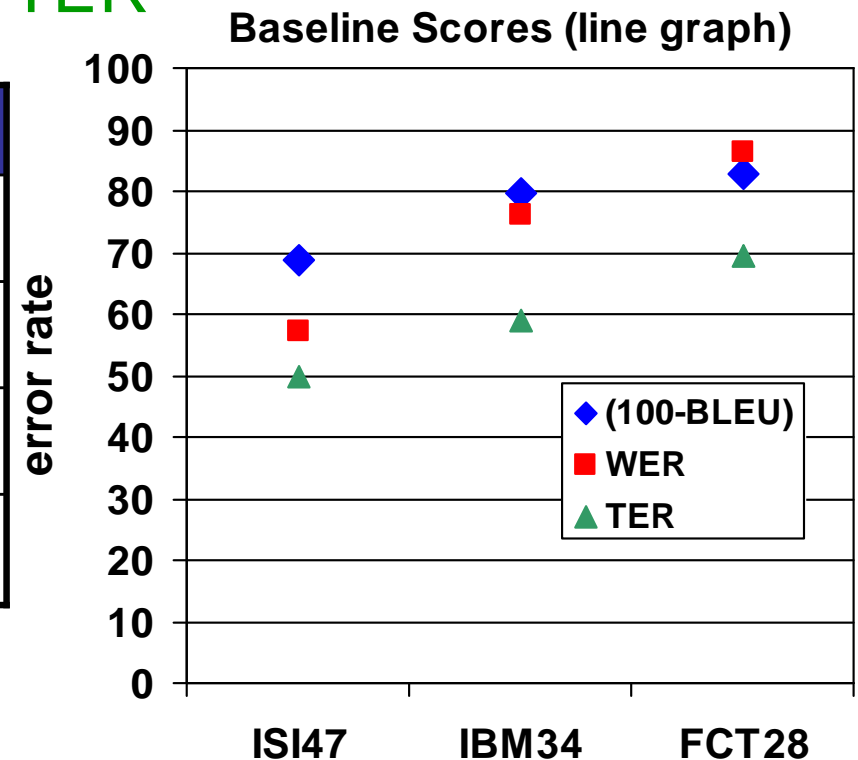
\* version, we created three orderings of the system/documents

No “hands-on edit” training for post editors

# Scoring - Baseline

- Original MT scored against gold-standard reference
  - All reported scores are case sensitive
  - We examine **BLEU**, **WER** and **TER**

Baseline Scores			
	ISI	IBM	FCT
(100-BLEU)	68.8%	79.5%	82.7%
WER	57.3%	76.2%	86.2%
TER	49.7%	58.8%	69.5%

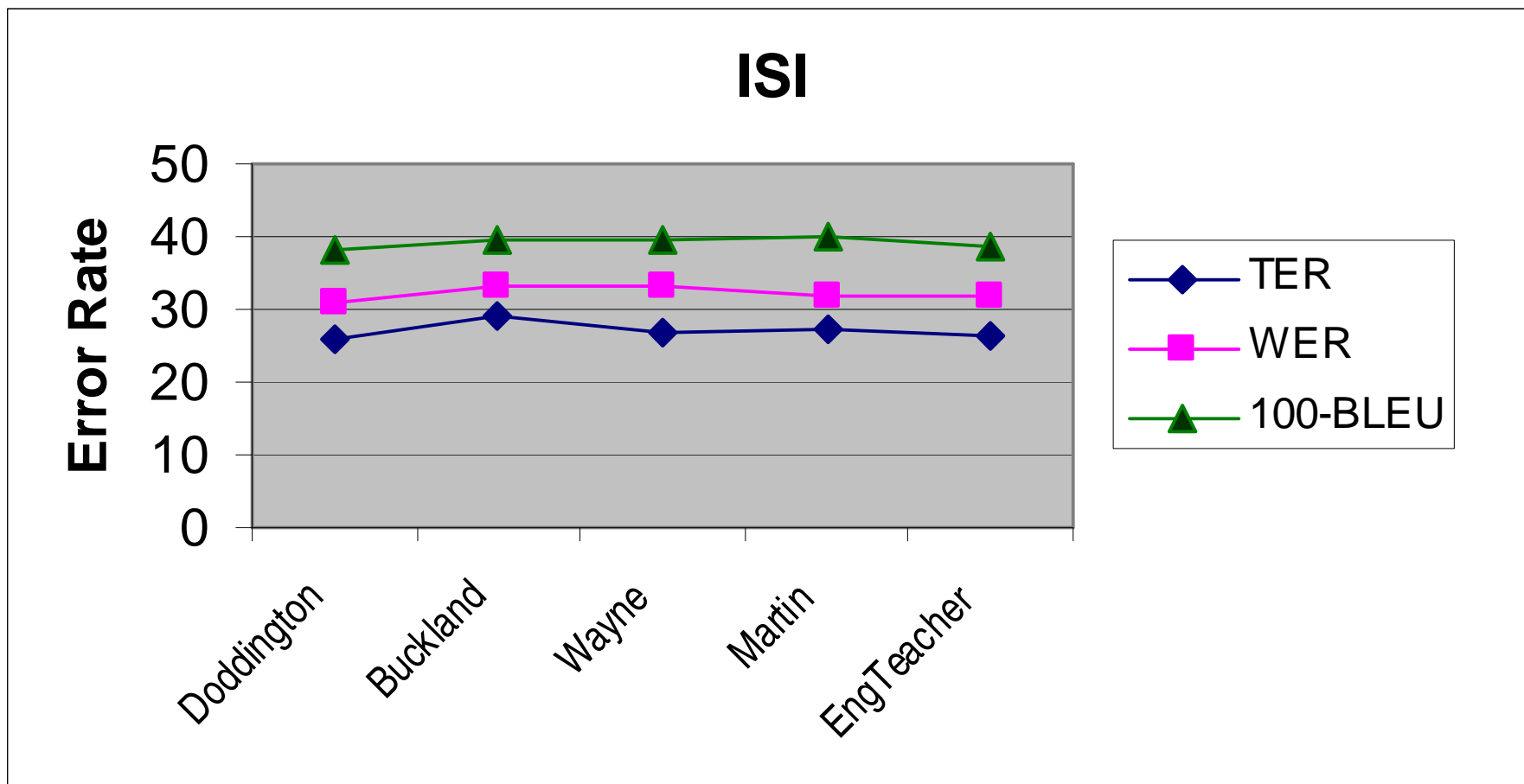


System ranking does NOT change with different error metrics



# Scoring -ISI

- Edited MT scored against Original MT



# Post Editor Agreement

- All three POC exercises will measure post editor agreement
  - We will measure the similarity of the *number of edits* each post editor makes in order to capture the complete meaning of the gold standard reference
- We use three automatic measures:
  - BLEU
  - WER (SCLITE)
  - TER (BBN introduced at MT05 workshop)
- We're limiting our PE agreement analysis to the ISI system

# PE Agreement: Document Order

- For each editor, we assign a rank from 1-10 to each document based on the document BLEU score
  - The average across the 5 editors establishes an average rank per document
    - Repeat for WER and TER metrics
- For our analysis we order the documents by the average document rank across the three metrics (*example follows*)

# PE Agreement in Rank (ISI, BLEU)

Sorted by:

MT vs GS (orig. %ile)	Document	Doddington	Buckland	Wayne	Martin	EngTeacher	AVG
1 (95)	XIN20040106.0051	1	2	1	2	1	1.4
2 (65)	XIN20040115.0212	2	4	2	1	2	2.2
4 (75)	XIN20040118.0127	3	1	4	4	3	3
5 (35)	XIN20040114.0251	4	6	3	3	4	4
3 (85)	AFA20040103.5700	5	3	5	6	6	5
7 (45)	AFA20040105.7700	6	8	8	5	7	6.8
8 (55)	AFA20040103.7710	8	5	7	8	8	7.2
9 (15)	AFA20040105.6200	7	7	6	7	5	7.8
6 (25)	AFA20040101.5100	9	10	9	9	9	9.2
10 (5)	XIN20040115.0055	10	9	10	10	10	9.8

 Colored cells match average ranking

# PE Agreement in Rank (ISI, **WER**)

Sorted by:


Document	Doddington	Buckland	Wayne	Martin	EngTeacher	AVG
XIN20040106.0051	1	2	1	2	1	1.4
XIN20040115.0212	2	3	2	1	2	2
XIN20040118.0127	3	1	4	4	3	3
XIN20040114.0251	4	6	3	3	6	4.4
AFA20040103.5700	5	4	9	7	4	5.8
AFA20040105.6200	7	7	6	5	5	6
AFA20040103.7710	8	5	5	8	7	6.6
AFA20040105.7700	6	8	8	6	8	7.2
AFA20040101.5100	9	10	7	9	9	8.8
XIN20040115.0055	10	9	10	10	10	9.8

 Colored cells match average ranking

# PE Agreement in Rank (ISI, TER)

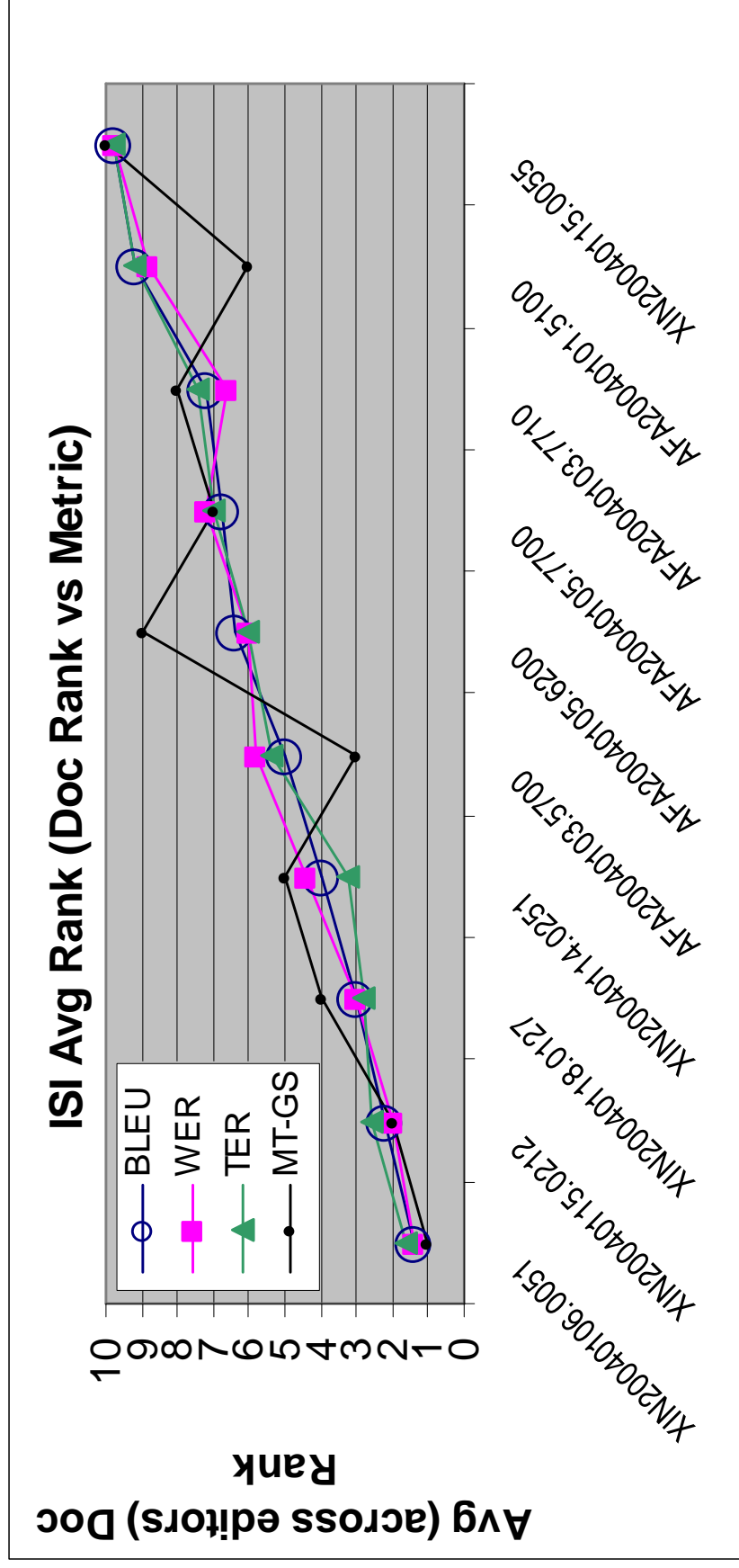
Sorted by:

MT vs GS (orig %ile)	Document	Doddington	Buckland	Wayne	Martin	EngTeacher	AVG
1 (95)	XIN20040106.0051	1	2	1	3	1	1.6
2 (65)	XIN20040115.0212	4	3	3	1	2	2.6
4 (75)	XIN20040118.0127	2	1	4	4	3	2.8
5 (35)	XIN20040114.0251	3	5	2	2	4	3.2
3 (85)	AFA20040103.5700	5	4	5	7	6	5.4
7 (15)	AFA20040105.6200	7	7	6	5	5	6
9 (45)	AFA20040105.7700	6	8	8	6	7	7
6 (55)	AFA20040103.7710	8	6	7	8	8	7.4
8 (25)	AFA20040101.5100	9	10	9	9	9	9.2
10 (5)	XIN20040115.0055	10	9	10	10	10	9.8

 Colored cells match average ranking

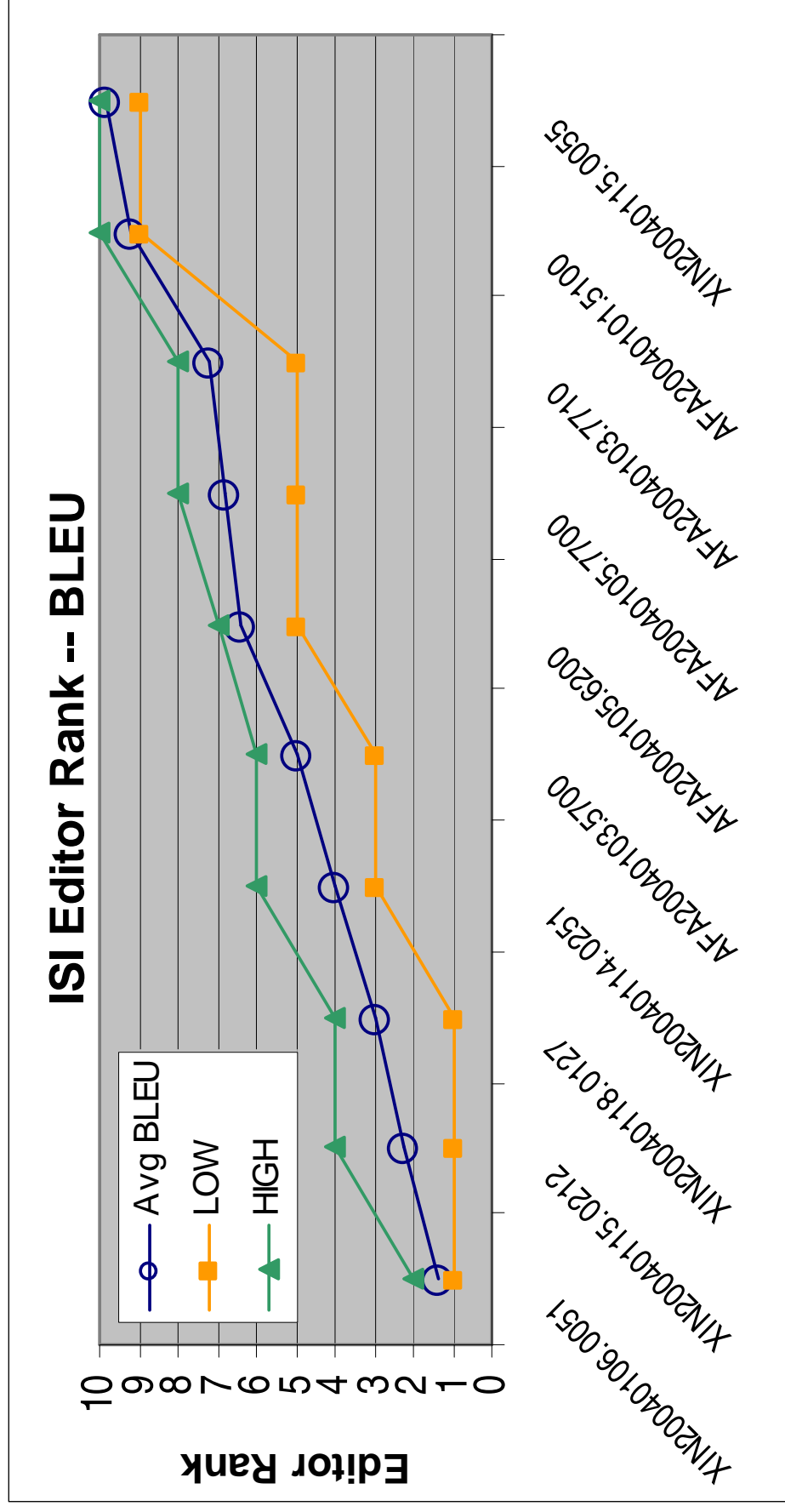
# PE Agreement

- Average document ranks are reasonably consistent across metrics



# PE Agreement by Rank for BLEU

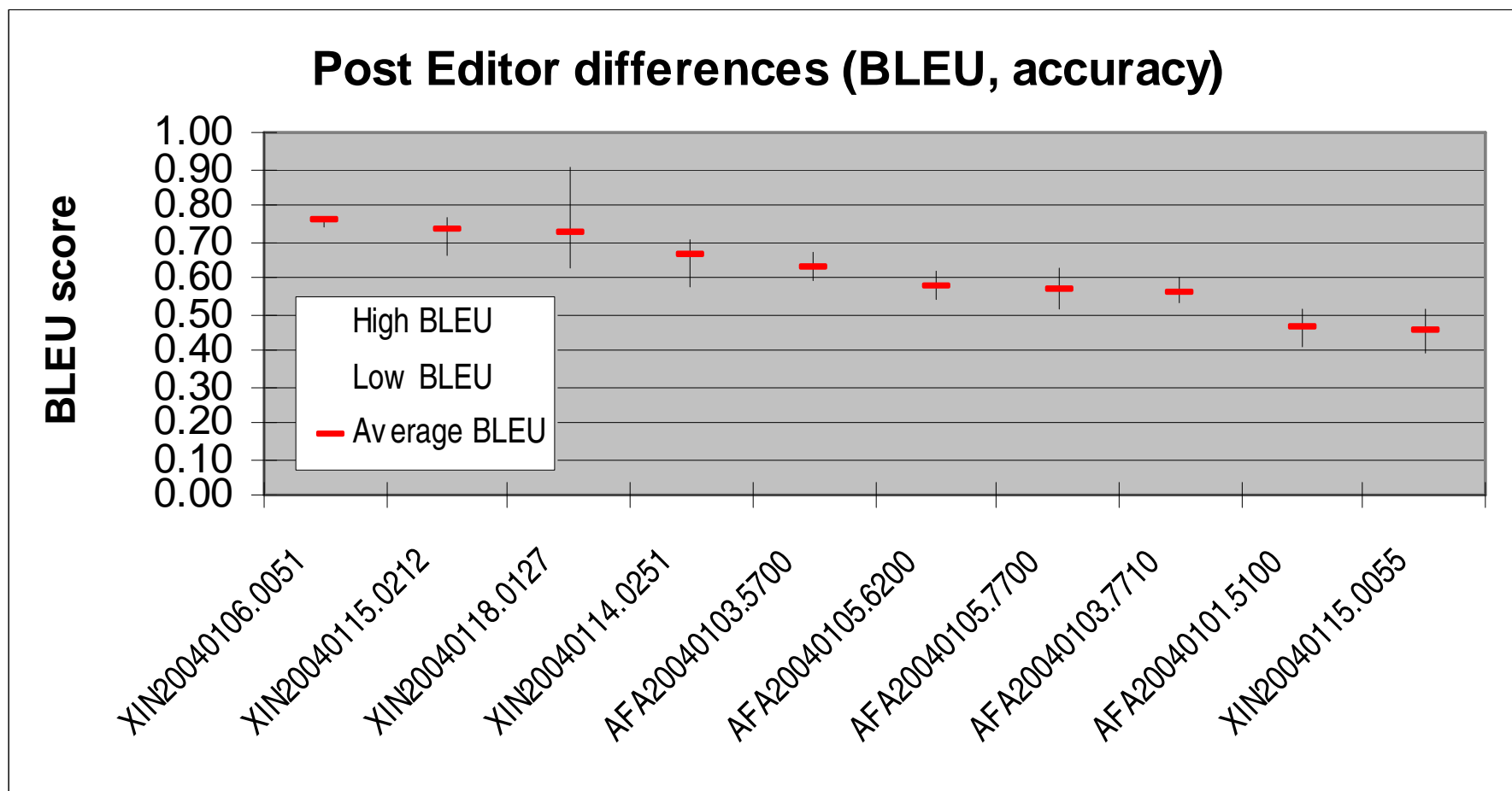
- Variation amongst PE's
  - Difference of three positions appears common





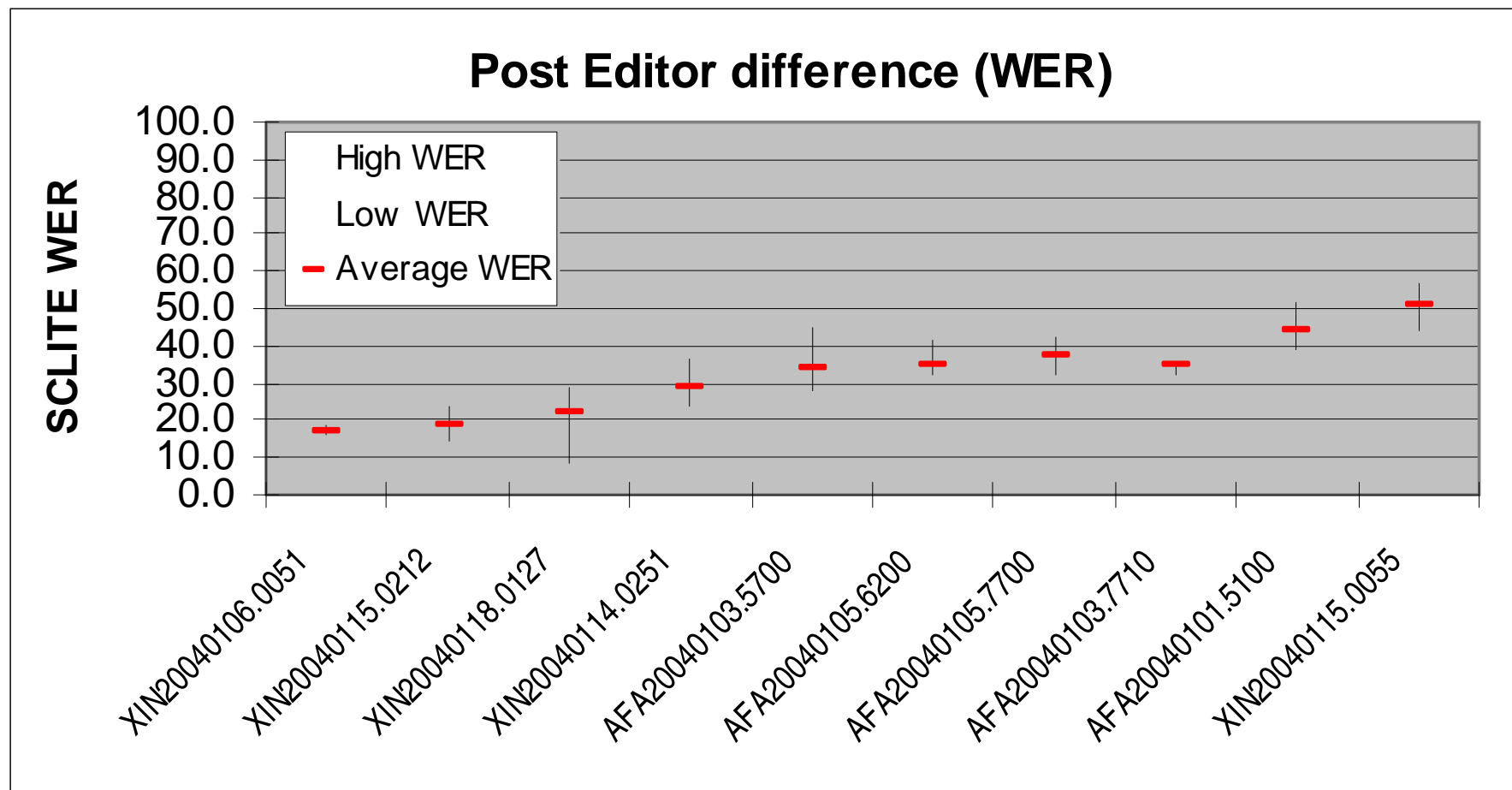
# Post Editor Agreement by BLEU Score

- We observe one obvious outlier among the 5 editors and the 10 documents
  - Average ratio (high / low) = 1.2 (*median: 1.2*)



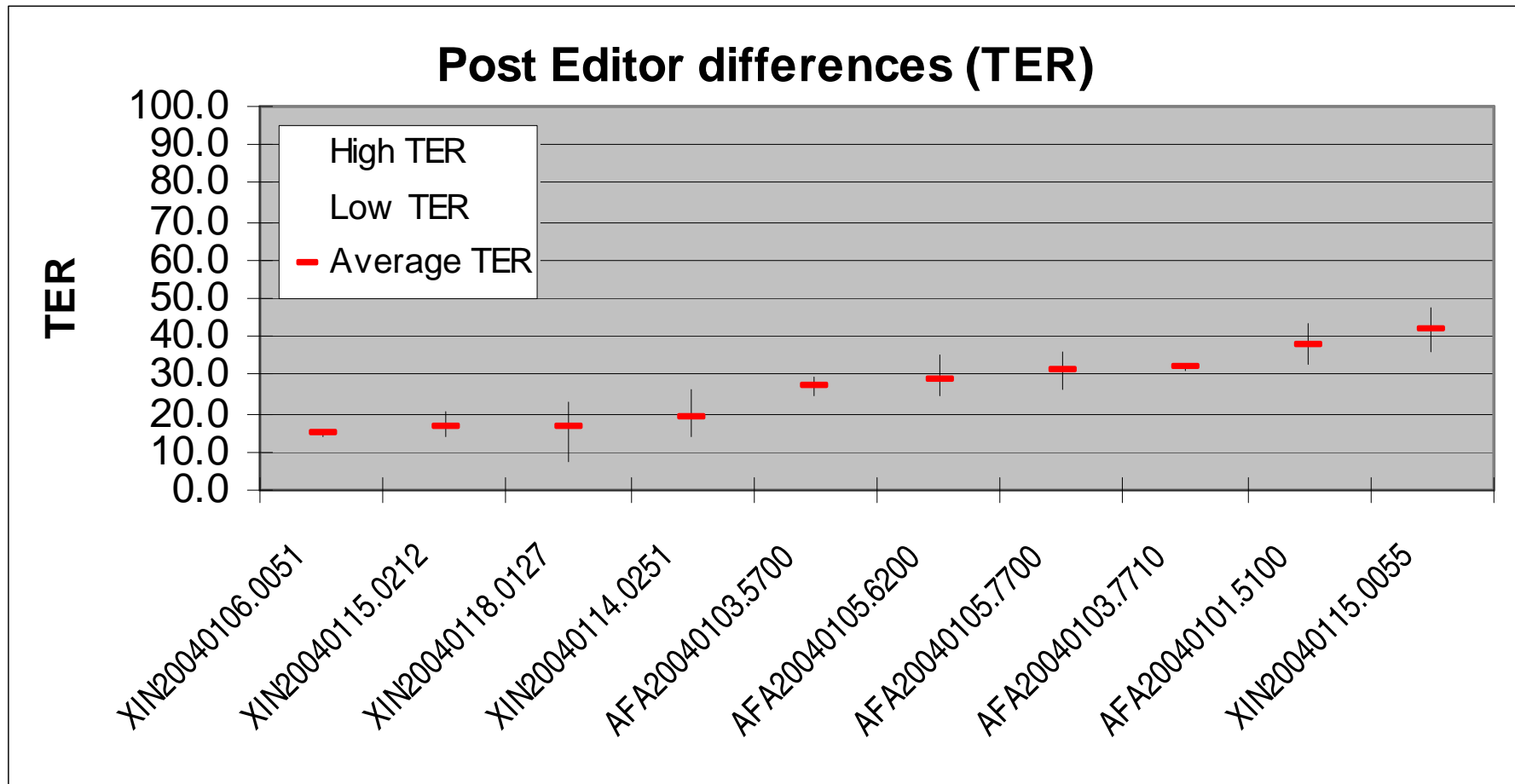
# Post Editor Agreement for WER

- We see similar variation
  - Average ratio (high / low) = 1.58 (*median: 1.34*)



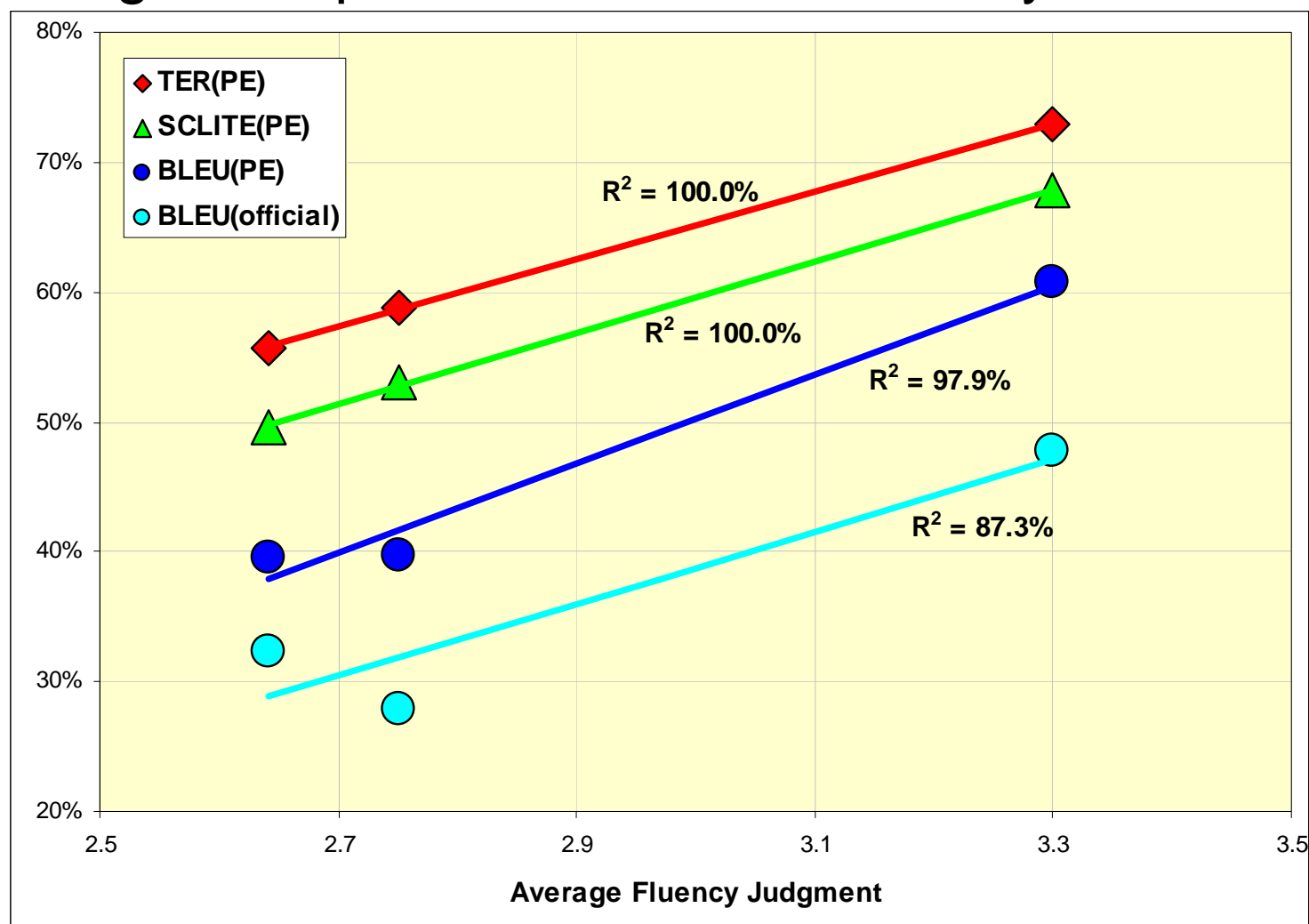
# Post Editor Agreement for TER

- We see similar variation
  - Average ratio (high / low) = 1.54 (*median: 1.34*)



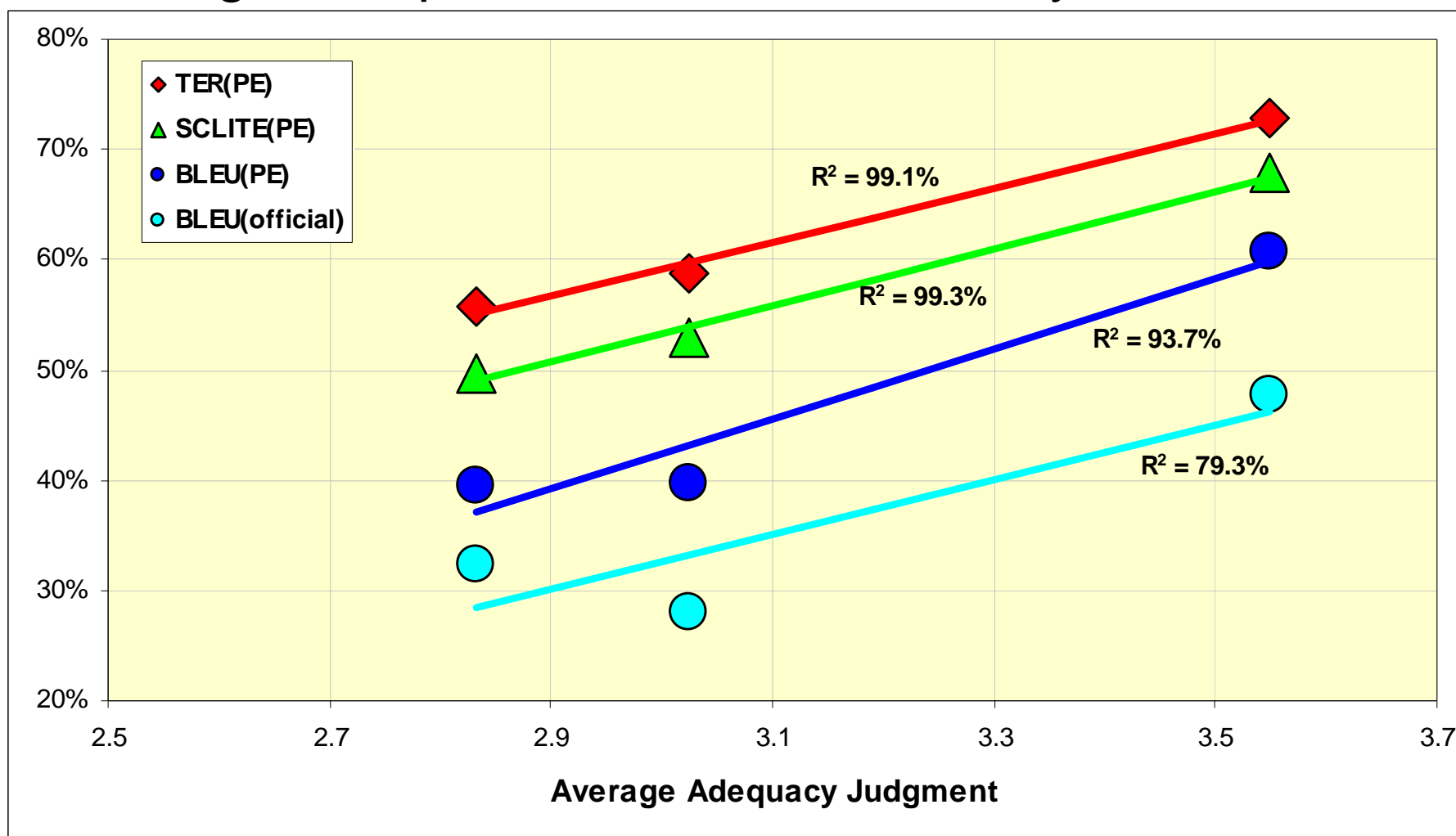
# Correlation with Human Assessments

- Metrics vs. Fluency
  - Aggregated over all editors
  - 2 segments per doc were assessed by humans



# Correlation with Human Assessments

- Metrics vs. Adequacy
  - Aggregated over all editors
  - 2 segments per doc were assessed by humans



# Test Set Size

- Goal
  - 95% confidence in differentiating absolute performance differences of 5% in TER
- How big must the test set be to achieve this?
- We can use the observed variances to form a mathematical model of required test set size for finding differences in relative system performance at a pre-defined significance level
  - Technique was successfully applied to EARS CTS “progress” set for test set size determination
- Initial findings (work in progress) – updated from July 7 talk
  - FOR SYSTEMS OPERATING AT ABOUT THE ISI LEVEL OF PERFORMANCE
  - ISI sample standard deviation is 12-15% (BLEU or TER)
  - 30 docs may be enough for 95% confidence measure for  $\pm 5\%$  absolute in system difference.
    - 150 docs for  $\pm 2\%$
    - 600 docs for  $\pm 1\%$
  - at 90% confidence measure early estimates indicate:
    - 20 docs for 5%
    - 100 docs for 2%
    - 200 docs for 1%
  - Further study is required

# Conclusions (1)

- We can expect post editors to work at a rate of 200-250 words per hour
  - possibly faster for text input as they become proficient
  - possibly slower with speech input
- We would like to see the ratio between edited MT high and low scores (regardless of the metric) move closer to one
  - BUT, current editing scores correlates well with human assessments
- Protocols needed for creating Gold Standard reference data
  - majority rules technique adopted due to time constraints
- Variance for document scores were higher for ISI (the better performing system)
  - One post editor notes, ISI was probably being edited, while the other two system outputs were just a starting point in rewriting the reference.

# Conclusions (2)

- Short (easy to remember) guidelines are good
- We must continue to refined the guidelines to promote inter-editor agreement

## Suggested Modifications

1. Make the MT output have **the same meaning** as the human translation.
2. Make the MT output **understandable** and **fluent English**
3. Use only necessary punctuation to satisfy the above criteria. Sentence-like units must have sentence-ending punctuation. Do not insert, delete, or change other punctuation merely to follow optional traditional rules about what is “strictly correct.”
4. If words/phrases/punctuation in the MT output or the reference human translation are completely acceptable, don’t take editorial license by adding/substituting new words.
5. Dates, as well as the commas and decimal points in numbers, should be formatted according to U.S. conventions (for example, convert 23-2-2004 to 2-23-2004).

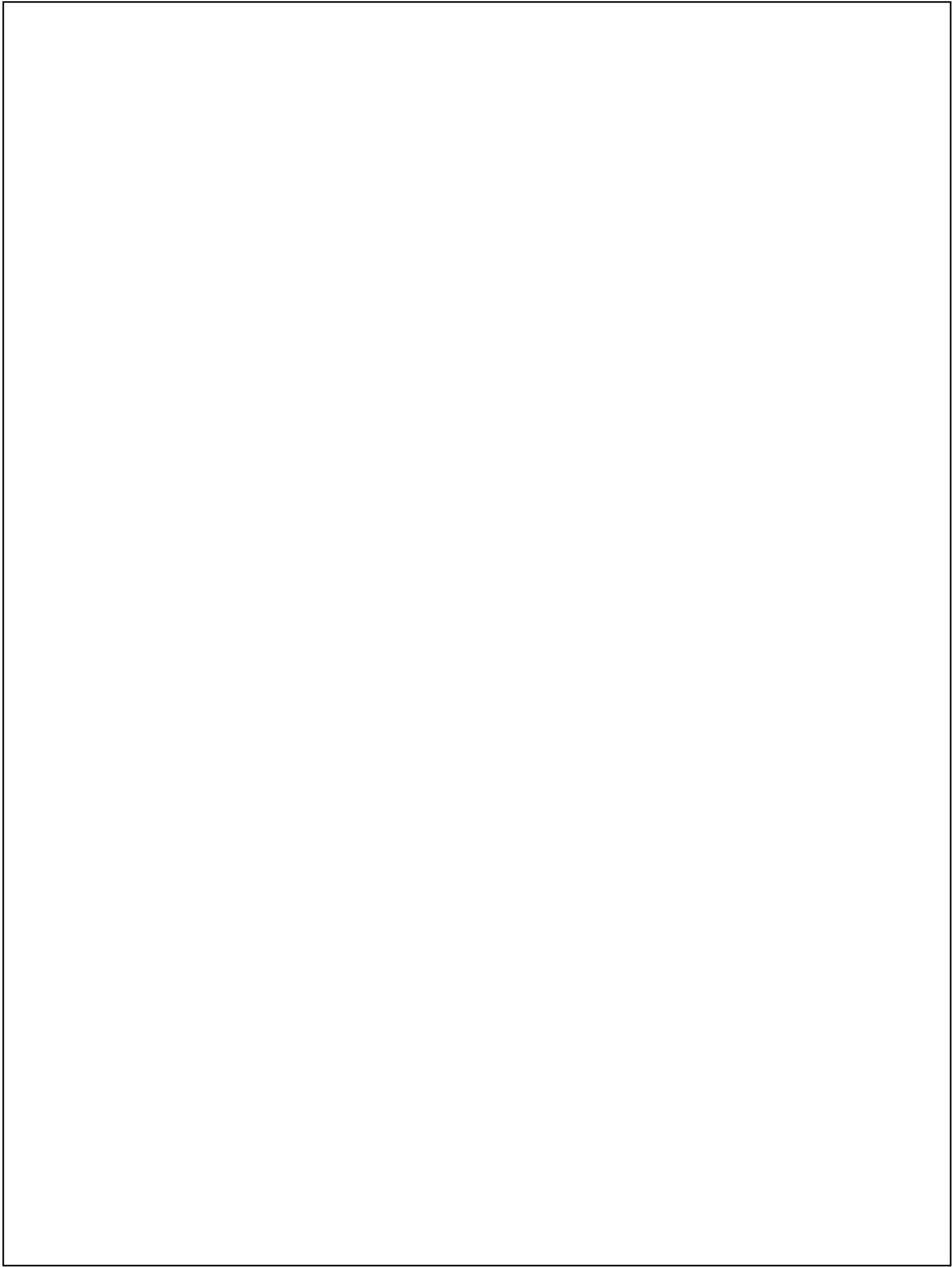


# Conclusions (3) What we learned

- Now ready to begin POC2 (Chinese text)
  - We will consider using MT05 data
    - Better system performance
    - Human assessments available for complete documents
- Will follow similar protocols as used for POC1, but will make modifications based on what we learned:
  - Updated guidelines
    - clearer set of 5 rules
    - stress how to handle punctuation and dates
    - use actual examples
  - Updated editing tool – in response to editor's comments
- New post editors will complete a well-defined training session before editing begins
  - First three documents won't be part of the exercise
- Will consider using 20-25 documents for 2 systems
- Will establish tighter guidelines for creating the gold-reference
- Begin to establish a GLM file for preprocessing MT to edit

# Preparing for the POC3 exercise

- Speech input for Arabic
- Data domain will be broadcast news so we can leverage against existing resources
- We have
  - BN audio
  - A single source language reference
  - Source language ASR
- We need help
  - MT output run on source ASR and source reference
  - English translation(s) of foreign BN audio



# Post Editor's comments

- NIST-1
  - Found the process painful. Notes instances of poor English syntax and likely errors in the reference. Domain knowledge is important along with language ability for good translations.
  - I may not have been consistent in handling punctuation, capitalization, and the format required for proper names and expressions. (PE's may require more explicit guidelines)
  - There is room for judgment as to what constitutes same meaning. There are issues of verb tenses and use of pronouns, and how finely tuned the English syntax has to be to be acceptable.

# Post Editor's comments

- High school teacher
  - The biggest challenge of editing the machine translation text, aside from the minimal time given, was having to accept phrasing that was not ordinary-sounding English. I had to constantly ask myself, “Does it make sense and does it carry the same meaning as the reference?” If it did, then I forced myself not to change the wording to make it sound better or easier to read. The end result is that some of the segments are not fluent English, but understandable just the same.
  - Keeping the guidelines in mind, I was sometimes confused as to whether or not to correct some grammatical errors, for example: commas surrounding appositives, correct spelling or capitalization of proper nouns and acronyms, dates and time in military language, and the syntax or arrangement of words within a sentence. I tended to make the changes only when meaning was disrupted due to mistakes of this nature

# Post Editor's comments

- NIST-2
  - The guidelines were good, easy to follow.
  - I had a few cases of text in quotes. Since the text was in quotes, but not proper English, I wasn't sure if I should edit it (I believe I did).
  - I like the interface, it was easy and pleasant to work with, but, the text was too small
  - I enjoyed the task. I believe I would have preferred to work on it 2 or 3 hours a day for a week rather than complete it all at once.